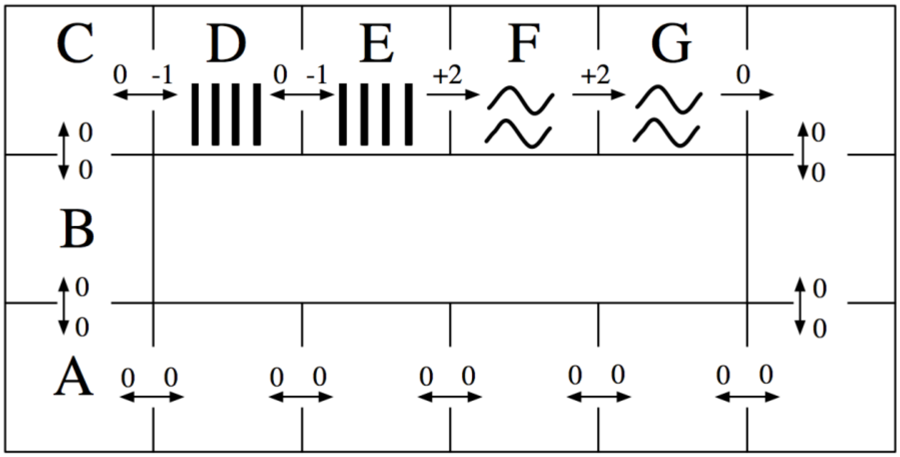
CSE 415 Computer Science and Engineering, University of Washington

Name: Travis Xie

Waterpark World
Part II: Q-Learning

Here is the Waterpark World MDP with seven of its 14 states labeled.



Fill in the blank cells of the following table with the Q-values that result from applying the Q-update for the transition specified on each row. You may leave blank those Q-values that are unaffected by the current update. Use discount γ = 1.0 and learning rate α = 0.5. Assume all Q-values are initialized to 0. (Note: the specified transitions would not arise from a single episode.) For your convenience, the Q-update formula is provided beneath the table below.

	$Q(D, \text{left})$	$Q(D, \text{right})$	$Q(E, \text{left})$	$Q(E, \text{right})$
Initial:	0	0	0	0
Transition 1: $(s = D, a = \text{right}, r = -1, s' = E)$		-0.5		
Transition 2: $(s = E, a = \text{right}, r = +2, s' = F)$				1
Transition 3: $(s = E, a = \text{left}, r = 0, s' = D)$			0	
Transition 4: $(s = D, a = \text{right}, r = -1, s' = E)$		-0.25		

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + (\alpha)\left[r + \gamma \max_{a'} Q(s',a')\right]$$

(V-ST-17-05-19. This exercise is based on an example used at U.C. Berkeley)